Distance-based Kriging relying on proxy simulations for inverse conditioning

David Ginsbourger, Bastien Rosspopoff, Guillaume Pirot, Nicolas Durrande, Philippe Renard

Distance-based Kriging relying on proxy simulations for inverse conditioning

David Ginsbourger[a], Bastien Rosspopoff[b], Guillaume Pirot[c], Nicolas Durrande[d], Philippe Renard[c]

[a]*Department of Mathematics and Statistics, University of Bern*
Alpeneggstrasse, 22, CH-3012 Bern, Switzerland
[b]*Ecole Nationale Supérieure des Mines*
FAYOL-EMSE, LSTI, F-42023 Saint-Etienne, France
[c]*Centre of Hydrogeology and Geothermics, University of Neuchâtel*
11 Rue Emile Argand, CH-2000 Neuchâtel, Switzerland
[d]*Sheffield Institute for Translational Neuroscience*
385a Glossop Road, Sheffield S10 2HQ, UK

Abstract

We consider the problem of rapidly identifying, among a large set of candidate parameter fields, a subset of candidates whose responses computed by accurate forward flow and transport simulation match a reference response curve. In order to keep the number of calls to the flow simulator computationally tractable, a recent distance-based approach relying on fast proxy simulations is revisited, and turned into a non-stationary Kriging method. The covariance kernel is obtained by combining a classical kernel with the proxy function, hence generalizing the idea of random field deformation to high-dimensional Computer Experiments. Once the accurate simulator has been run for an initial subset of models and a Kriging metamodel has been inferred, the predictive distributions of misfits for the remaining geological models can be used as a guide to solve the inverse problem in a sequential way. The proposed algorithm, *Proxy-based Kriging for Sequential Inversion* (PROKSI), relies indeed on a variant of the *Expected Improvement*, a popular criterion for Kriging-based global optimization. A statistical benchmark of ProKSI's performances finally illustrates the efficiency and the robustness of the approach when using different kinds of proxies.

Email addresses: ginsbourger@stat.unibe.ch (David Ginsbourger), brosspopoff@etu.emse.fr (Bastien Rosspopoff), guillaume.pirot@unine.ch (Guillaume Pirot), n.durrande@sheffield.ac.uk (Nicolas Durrande), philippe.renard@unine.ch (Philippe Renard)

Preprint submitted to Advances in Water Resources *May 16, 2012*

1. Introduction

Inverse techniques are one of the corner stones of groundwater modeling. In broad terms, their aim is to identify model structure and model parameter values from observed state variables. In practice, a wide range of approaches have been developed [1, 2, 3, 4, 5]. Most often, the inverse problem is formulated in a least-square manner. A data misfit quantifies the difference between measured and calculated state variables, it is a function of the unknown parameter values. The aim is then to find models minimizing the misfit. To avoid unrealistic parameter sets, various model regularization schemes can be employed. Less frequently, the problem is solved in a Bayesian framework, and instead of providing a single unique solution (the best estimate), the aim is to recover the posterior probability distribution of the model parameters knowing the values of the state variables. When the problem is non-linear and when the prior distributions for the parameter fields are not Gaussian, it is generally not possible to provide an explicit analytical expression of the posterior distribution. In such situations, one must rely on computational resources and statistical sampling techniques [6, 3, 7, 8] to get a representative sample (ensemble of parameter fields) of the posterior distribution. The most advanced techniques are based on Markov Chain Monte Carlo (MCMC) approaches [9, 10, 11, 8]. They consist in generating samples from the prior distribution of parameters and running the forward flow and transport model on those samples to evaluate the misfit and consequently the likelihood of each particular sample (by comparing the computed state variables with the actual measurements) before accepting the sample or not in the posterior ensemble. The practical difficulty involved with MCMC is that the calculation of the likelihood function is often computationally very demanding and this inhibits the user to let the MCMC algorithm run for a sufficiently large number of iterations to enable convergence [12, 11]. Similar computational issues arise in optimization problems related to groundwater management: if each evaluation of the objective function that has to be minimized requires a significant amount of computational resources, it may become infeasible to reach the optimum in a reasonable time and special techniques must be developed.

To reduce the computational demand, one can use the concept of response surface, or *metamodel*. The response (misfit or objective function) of the flow simulator is computed for a small set of candidate parameter fiels and predicted by the metamodel in the remaining part of the parameter space. Various interpolation techniques can be employed such as radial basis functions, splines, or kriging [13, 14, 15, 16, 17, 18]. The main advantage of using kriging is its ability to provide both a prediction of the possible response (kriging mean m) and a corresponding prediction uncertainty (kriging variance s^2). The prediction uncertainty drops to zero where the response has actually been computed with the numerical model and increases when moving away from those points. If we consider a global optimization problem consisting in finding parameter values minimizing the model response, one can use m and s^2 to express a trade-off between the exploitation of the response function (finding locations where the estimated values m are low) and exploration of the design space (finding locations where the prediction is the most uncertain). Combining these two ideas gave birth to the *Expected Improvement* (EI) criterion [19]: at every location (within the parameter space), the kriged response surface is used to estimate the expected value of the possible improvement (difference between the possible value at that location and the value of the current minimum obtained with the numerical model). The value of the input parameter vector with the highest EI is then chosen to run the numerical model again and update the response surface. Such approaches based on kriging metamodels have been very successfully used for sequential design of computer experiments since the development of the *Efficient Global Optimization* algorithm [20] in the late 1990's. Several other criteria were later proposed (see [21] for an overview).

Another approach to reduce the computational demand is to use a concept of distance between parameter fields [22, 23, 7]. Several types of distances can be defined, but the important point is that the distance should be defined such that it can be computed rapidly and used as a guide to predict if two parameter fields will lead to similar or different responses. For example, Suzuki et al. [24] used the Hausdorff distance to quantify the differences in the geometry of complex 3D models (having different fault systems, horizon geometries, etc.), coupled with the neighborhood algorithm [25] to search efficiently, within the prior ensemble,

the models that match field observations of oil production. Scheidt and Caers [23] propose a general framework based on the concept of distance to quantify uncertainty. In their example, the problem consists in estimating oil recovery in a production well. The models all have the same geometry, but very different parameter fields (obtained using multiple-point statistics with different training images). The prior ensemble is large and the aim is to obtain rapidly a good estimation of the uncertainty on the forecast. For that purpose, Scheidt and Caers [23] define the square distance between two parameter fields as the integrated square difference between the responses computed for the two models with a fast streamline solver. The distances between every pair of models is computed and used as the base for mapping all the models in an abstract metric space in which it is possible to select a small number of parameter fields covering comprehensively the variability of the complete ensemble. Running the forward two-phase flow numerical simulator only on this small number of selected models allows a fast and rather accurate estimation of the uncertainty. Going a step further, Caers et al. [26] use the same framework to formulate the inverse problem. They propose to solve a so called pre-image and post-image problems which consist in generating parameter fields which are located at a pre-specified location in the metric space corresponding to the solution of the inverse problem.

A final direction that seems promising to reduce the computational demand is the joint use of a pair of complex and simple models [27, 28, 29]. The distinction between the complex and simple models is not straightforward, but to remain general we can state that the complex model tends to account for all the important and relevant physical processes as well as all the necessary geometrical complexity of the reservoir. On the opposite, the simple model neglects some aspects of this complexity with the aim of being much more computationally efficient. The simplification may be based on neglecting some physical processes, it may be based on reducing the space dimension of the problem (2D instead of 3D), it may also be based on a coarse spatial or temporal resolution. In the remaining of this paper, we will use the terminology *accurate* model for the complex one, and *proxy* for the simple one. To use a combination of accurate and proxy models in practice, one needs to establish a link between the two. Several approaches can be devised. For example,

Doherty and Christensen [29] identify some parameters of the proxy model by solving an inverse problem where the results of the accurate model have to be reproduced.

In this paper, we propose to link an accurate and a proxy model using a distance-based kriging metamodel. It allows to forecast the possible response of the accurate model as it is done with traditional kriging metamodels. However, those methods are usually limited to parameter spaces of small dimensions. This makes their application for the identification of complete parameter fields impossible. The novelty of the proposed approach lies therefore in the way we define the covariance kernel at the core of the kriging metamodel. The concept is simple, we assume that the same parameter fields can be used as input data for the proxy and the accurate model. As suggested by Caers and his collaborators [22, 23, 26, 7] we use the distance in proxy responses, but we include that distance into the covariance kernel of the kriging equations. The consequence is a drastic reduction of the dimensions of the problem allowing to infer the statistical parameters of the covariance. Once the statistical relation between the proxy and the accurate model is established, it can be used to predict the accurate response and its uncertainty for any model whose proxy response is known. It can also be updated when new runs of the accurate model become available. This general idea can be applied for a very wide range of problems.

The main aim of this paper is therefore to describe the concept of the distance-based kriging technique. We also illustrate how this technique can be used in a sequential algorithm aiming at quickly identifying a set of parameter fields whose responses computed with an accurate model match some reference data. Because the purpose, in an inverse problem, should not only be to find the global minimizer(s) but more to sample from a posterior distribution, we propose a variant of the EI criterion meant to spend more time exploring the possible various minima of the misfit function than EI.

For illustration purpose, we consider a simple flow and solute transport problem. The geological heterogeneity is modeled using a multiple-point statistics technique [30] allowing to account for prior geological knowledge typical for a fluvio glacial environment. Numerous experiments with a randomization procedure are conducted to test the robustness of the method.

The paper is organized as follows. In section 2 we first give an overview of the sequential algorithm used to solve the inverse problem. Then we describe in detail the proposed kriging metamodel in section 3. The equations of ordinary kriging are recalled, with a focus on the crucial role of the covariance kernel. The original kernel underlying our work is introduced, followed by a discussion on its interpretation as well as its mathematical foundations. To close the section, we give practical details concerning the estimation of covariance parameters. We then end the presentation of the method in section 4 by describing how the sequential search is driven. Section 5 and 6 are dedicated to results and discussion. We first introduce a case study to illustrate the methodology. Then we present the obtained experimental results and statistically assess the performance of the method based on a benchmark of 100 randomly chosen reference curves. We finally conclude and propose a few theoretical and practical perspectives in section 7.

2. Overview of the sequential algorithm

The proposed sequential algorithm is named *Proxy-based Kriging for Sequential Inversion* (ProKSI). Its aim is to identify rapidly, within a large ensemble of parameter fields, the ones whose responses computed with the accurate model fit some reference curve. In practice, the algorithm consists in sequentially selecting among all the available models which one will be used as input for the accurate numerical model at the next iteration (Fig. 1 to 2). Before sketching the key phases of the algorithm, let us set a few notations.

Each candidate parameter field is denoted $\mathbf{x}_i \in E$ $(1 \leq i \leq N)$, where E is a vector space, typically of dimension 10^4 to 10^6 when representing a discretization of the subsurface. In the following examples, $\mathbf{x}_i$ represents a categorical field obtained from multiple-point statistics simulation. But the proposed methodology is more general and can be applied without much modifications to models having various geometries or even based on different conceptual assumptions. The only requirement is that it is possible to compute the accurate and proxy responses for any of those input models.

The accurate numerical simulator is considered as a function f returning a vector of values. In the example, we assume more specifically that f returns a breakthrough curve

(concentration versus time):

$$f_{\mathbf{x}} : t \in [0, T] \rightarrow f_{\mathbf{x}}(t) \in \mathbb{R}_+ \tag{1}$$

for any input $\mathbf{x} \in E$, where t represents the time. The space of such curves is denoted by F.

Now, given a reference curve $f_{\text{ref}} \in F$, the goal is to recover in a limited time which $\mathbf{x}_i$'s $(1 \leq i \leq N)$ minimize the misfit $g^\circ(\mathbf{x}) := d(f_{\text{ref}}, f_{\mathbf{x}})$, where d is some metric on F. For example, if we use the L^2 norm, the misfit will be expressed as:

$$g^\circ(\mathbf{x}) = \int_0^T \left(f_{\text{ref}}(t) - f_{\mathbf{x}}(t)\right)^2 dt \tag{2}$$

Ideally, one wishes to get a good picture of the subset of input fields leading to a good fit, relying on a fixed number of evaluations $k < N$ dictated by computation time constraints. In addition to the costly f, we assume that a "*proxy*" $p : E \rightarrow F$ is available, providing an approximate solution to the flow and transport equations significantly faster than f. Depending on the context, p may stem for instance from an auxiliary simulator solving similar equations with simplified physics, or from degrading the accurate simulator f by reducing the time or spatial resolution.

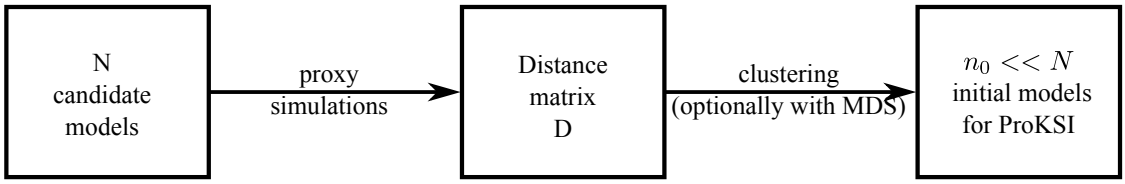

Figure 1: Initialization steps of the ProKSI algorithm.

The ProKSI algorithm starts with a series of initialization steps (Fig. 1):

1. A large number N of $\mathbf{x}_i$'s are generated (e.g., by multiple-points statistics simulation).
2. The proxy responses $p(\mathbf{x}_i, t)$ are computed for all $\mathbf{x}_i$'s $(1 \leq i \leq N)$. The distances $d_{i,j}$ between the proxy responses of any pair of input fields are then computed:

$$d_{i,j} = \int_0^T \left(p(\mathbf{x}_i, t) - p(\mathbf{x}_j, t)\right)^2 dt \tag{3}$$

This allows assembling the distance matrix D between all proxy responses.

3. A clustering technique (k-means) is used to group the models in n_0 classes. For each class, the models that are the closest to the centroïd are selected to get a subset $\mathbf{X}_{n_0} = \{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{n_0}}\}$ of n_0 initial models (See Fig. 6(a)). Multi Dimensional Scaling (MDS) is optionally used to map all the input parameter fields in a small-dimensional euclidean space (Fig. 6(a)).

For each of those n_0 models, the accurate response f_{i_j} is computed with the accurate numerical solver. We obtain a vector $\mathbf{g}^\circ = \{g^\circ_{i_1}, \ldots, g^\circ_{i_{n_0}}\}$ ($g^\circ_{i_j} := g^\circ(\mathbf{x}_{i_j})$, $1 \leq j \leq n_0$) containing the misfits for the n_0 models. The values of $\mathbf{g}^\circ$ are transformed using a power-law $g_{i_j} = \left[g^\circ_{i_j}\right]^a$ to obtain a sample $\mathbf{g}$ with a close-to-Gaussian distribution. Note that here and in the sequel, the value of a is obtained by minimizing the skewness of the sample of transformed values $\{g_{i_j}, 1 \leq j \leq n_0\}$.

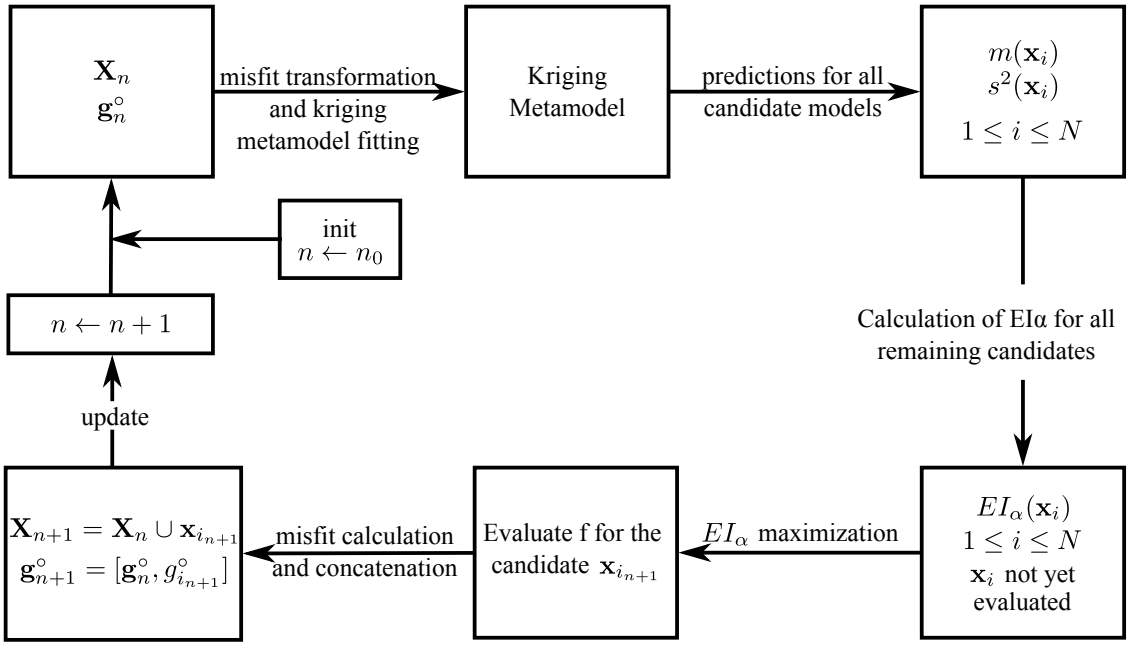

Figure 2: Sequential loop of the ProKSI algorithm.

A sequential loop (Fig. 2) then allows to select a new candidate model at each iteration on which to run the accurate solver. This enables building progressively a set of parameter fields with low misfit values. The steps in that loop are the following (n is first set to n_0):

1. If not already done, apply a normalizing transform to the sample of misfits (See detail

above). Estimate the covariance parameters τ, θ, and σ^2 by Maximum Likelihood as described in section 3. Compute the kriging mean $m(\mathbf{x}_i)$ and the variance $s^2(\mathbf{x}_i)$ for all inputs $\mathbf{x}_i \notin \mathbf{X}_n$.

2. After having computed the value of the modified expected improvement criterion $EI_\alpha(\mathbf{x}_i)$ (see section 4 for its definition) for all the remaining candidate models, Select a model with maximal EI_α value as next candidate, called $\mathbf{x}_{i_{n+1}}$.
3. Set $\mathbf{X}_{n+1} = \mathbf{X}_n \cup \{\mathbf{x}_{i_{n+1}}\}$. Compute $f_{\mathbf{x}_{i_{n+1}}}$ with the accurate numerical solver. Calculate the new corresponding misfit and append it to the vector of misfits: $\mathbf{g}^\circ_{n+1} = \{\mathbf{g}^\circ_n, g^\circ_{i_{n+1}}\}$. Go to step 1 and resume the search until a convergence criterion is met.

The algorithm stops when the EI_α reaches a prescribed lower threshold, or a desired number of evaluations has been done, for instance because the allocated search time is elapsed.

3. High-dimensional kriging with a proxy-based kernel

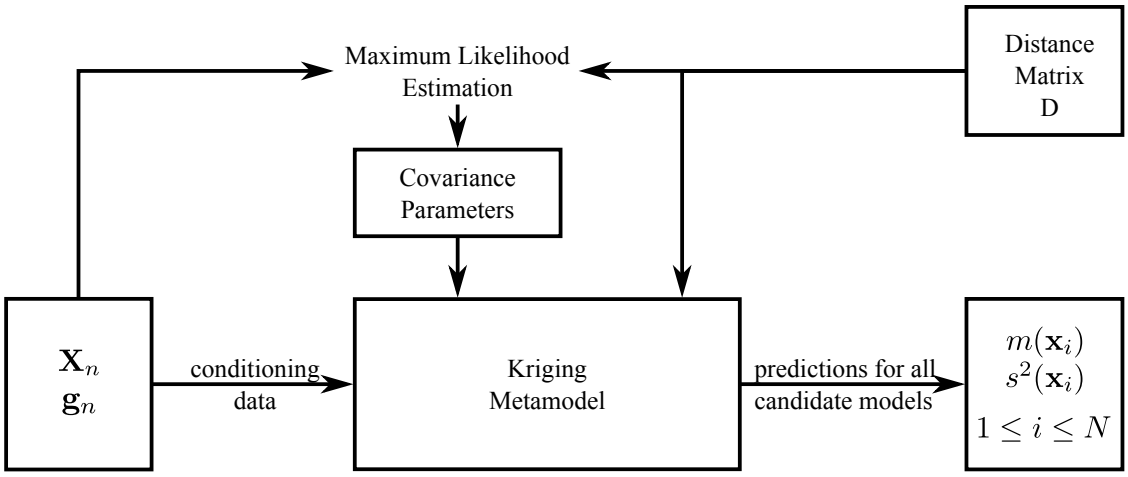

Figure 3: Overview of the main steps in proxy-based Kriging prediction (after misfit transformation).

The most important difference between the existing methods and what we propose here is the distance-based kriging approach. It lies at the heart of sequential algorithm described earlier in Figure 2. In this section, we will describe in detail how this step is performed. The main idea is to integrate the distance between proxy responses within the covariance kernel of the Kriging metamodel (Fig. 3).

3.1. Kriging for Computer Experiments

We adopt the framework of Gaussian Processes [17] to model the misfit between f_{ref} and the response of the accurate numerical model. The misfit g is assumed to be one realization of a Gaussian Process $(G_{\mathbf{x}})_{\mathbf{x}\in E}$, with mean function μ and covariance kernel k. We assume that μ is an unknown constant, as in the case of Ordinary Kriging. We denote $\mathbf{g}$ the vector of the known values of the misfit at the current design of experiments $\mathbf{X}_n := \{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_n}\}$ $(n \geq n_0)$, the Kriging mean $m(\mathbf{x}) = \mathbb{E}\left[G_{\mathbf{x}} | G_{\mathbf{x}_{i_1}} = g(\mathbf{x}_{i_1}), \ldots, G_{\mathbf{x}_{i_n}} = g(\mathbf{x}_{i_n})\right]$ and Kriging variance s^2 of the same random process at any arbitrary point $\mathbf{x} \in E$ write:

$$m(\mathbf{x}) = \widehat{\mu} + \mathbf{k}(\mathbf{x})^T K^{-1}(\mathbf{g} - \widehat{\mu}\mathbf{1}) \tag{4a}$$

$$s^2(\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{k}(\mathbf{x}) + \frac{(1 - \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{1})^2}{\mathbf{1}^T K^{-1}\mathbf{1}} \tag{4b}$$

where K is a $n \times n$ matrix with entries $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, referred to as the *covariance matrix of observations*, $\mathbf{k}(\mathbf{x}) := (k(\mathbf{x},\mathbf{x}_1), \ldots, k(\mathbf{x},\mathbf{x}_n))'$ is a $n \times 1$ *covariance vector*, and $\widehat{\mu} = \frac{\mathbf{1}^T K^{-1}\mathbf{g}}{\mathbf{1}^T K^{-1}\mathbf{1}}$ is the Best Linear Unbiased Estimator of μ.

One of the attracting features of Kriging is that m interpolates the observations (i.e. $\forall j \in \{1, \ldots, n\}$, $m(\mathbf{x}_{i_j}) = g(\mathbf{x}_{i_j})$). Furthermore, s^2 vanishes at the design points ($s^2(\mathbf{x}_{i_j}) = 0$), and gives a quantification of the prediction uncertainty at unobserved points. A very important feature is that both properties remain valid whatever the chosen covariance kernel k. Hence, equations (4a) and (4b) give a potentially infinite set of interpolating metamodels, and choosing a k adapted to the studied phenomenon appears to be a crucial issue in practice.

3.2. A new kernel for high-dimensional Kriging based on fast proxies

Designing a suitable covariance kernel over $E \times E$ is very challenging because E is a space of parameter fields of typical dimensions ranging between 10^4 to 10^6. Hence, taking kernels usually employed in d-dimensional ($d \approx 10$) cases, e.g., an anisotropic power exponential kernel, will a priori not make sense in the present framework. Alternatively, uncovering features of the models $\mathbf{x} \in E$ leading to similar response curves would be ideal.

Here, we take advantage of the proxy responses in order to define a relevant measure of

similarity. More precisely, we propose to use the following covariance kernel:

$$k(\mathbf{x},\mathbf{y}) := \sigma^2 \exp\left(-\frac{1}{\theta^2}\int_0^T (p(\mathbf{x},t)-p(\mathbf{y},t))^2 dt\right) + \tau^2 \mathbf{1}_{\mathbf{x}=\mathbf{y}} \tag{5}$$

In words, the closer two proxy curves associated with two parameter fields $\mathbf{x},\mathbf{y}$ are, the closer the fits to the reference are expected to be when running the accurate simulator with those inputs. In addition to this transformed Gaussian kernel, the term $\tau^2\mathbf{1}_{\mathbf{x}=\mathbf{y}}$ stands for the nugget effect, and allows to model a possible dissimilarity between the accurate responses of the inputs $\mathbf{x},\mathbf{y}$, even if their associated proxy responses are close or even identical.

In fact, the proposed covariance kernel k can be seen as a standard stationary Gaussian kernel over $F \times F$, chained with the "proxy operator", that is with the function p:

$$k(\mathbf{x},\mathbf{y}) := \sigma^2 \exp\left(-\frac{1}{\theta^2}||p(\mathbf{x})-p(\mathbf{y})||_F^2\right) + \tau^2 \mathbf{1}_{\mathbf{x}=\mathbf{y}} \tag{6}$$

where $||f||_F := \sqrt{\int_0^T f(t)^2 dt}$ ($f \in F$) stands for the L^2 norm over F (the functions of F being further assumed continuous). This basic fact ensures that the proposed kernel is an admissible covariance. k is indeed positive-semidefinite over $E \times E$ in virtue of the following property, for which a proof is proposed in appendix:

Property Let E and F be two arbitrary spaces. Given a positive-semidefinite kernel k_F over $F \times F$, the kernel k_E defined by

$$k_E(\mathbf{x},\mathbf{y}) := k_F(p(\mathbf{x}),p(\mathbf{y})) \tag{7}$$

is a positive-semidefinite kernel over $E \times E$ whatever the function $p : E \longrightarrow F$.

Note that in different contexts, similar methods relying on a change of variables within a positive-semidefinite kernel were already proposed, for example in [31] and subsequent works. Coming back to Eq. 5, the basis kernel k_F corresponding to Prop. 7 is none other than an isotropic Gaussian kernel $k_F(\mathbf{u},\mathbf{v}) = \sigma^2 \exp\left(-\frac{1}{\theta^2}||\mathbf{u}-\mathbf{v}||_F^2\right)$, parametrized by a sill σ^2 and a range parameter $\theta > 0$.

The next subsection focuses in detail on the chosen methodology for estimating the three parameters σ^2, θ, τ^2 from available data.

3.3. Parameter fitting for the proposed Kriging model

The approach chosen here for tuning the covariance parameters is Maximum Likelihood Estimation (MLE). MLE for covariance parameters in Ordinary Kriging settings relies on the assumption that $\mathbf{g}$ is one realization of a Gaussian vector with mean $\widehat{\mu}\mathbf{1}$ and covariance matrix K with entries driven by the parametric kernel k above. MLE then consists in maximizing the likelihood function for σ^2, θ, τ^2 given $\mathbf{g}$, or equivalently in minimizing:

$$l(\sigma^2, \theta, \tau^2; \mathbf{g}) := \log(\det(K)) + (\mathbf{g} - \widehat{\mu}\mathbf{1})^T K^{-1} (\mathbf{g} - \widehat{\mu}\mathbf{1}), \tag{8}$$

where K and $\widehat{\mu}$ are functions of $(\sigma^2, \theta, \tau^2)$. When $\tau^2 = 0$, it is known [32] that $\widehat{\mu} = \frac{\mathbf{1}^T R(\theta)^{-1} \mathbf{g}}{\mathbf{1}^T R(\theta)^{-1} \mathbf{1}}$, and the optimal value of σ^2 can be expressed as a function of θ only:

$$\sigma^{2*}(\theta) := \frac{1}{N} (\mathbf{g} - \widehat{\mu}\mathbf{1})^T R(\theta)^{-1} (\mathbf{g} - \widehat{\mu}\mathbf{1}), \tag{9}$$

where $R(\theta) := \frac{1}{\sigma^2} K(\sigma^2, \theta, 0)$ is the correlation matrix of $G_{\mathbf{X}_n}$. Minimizing l is equivalent to the one-dimensional minimization over θ of the so-called *concentrated log-likelihood*:

$$l_c(\theta; \mathbf{g}) := l(\sigma^{2*}(\theta), \theta, 0; \mathbf{g}). \tag{10}$$

When $\tau^2 > 0$, Eq. (9) is unfortunately no longer valid. In that case, a rigorous option would be to minimize l with respect to σ^2, θ, τ^2. However, when τ is very close to 0 as is often the case in practice (at least in the examples that we have investigated), it would be frustrating to throw up eq. (9) and loose the benefit of reducing the problem dimensionality to one because of tiny changes in the likelihood. Here we approach the problem sequentially, and preserve the concentration step at the price of a minor approximation. First, an estimate of τ^2 is derived based on variographic considerations. Then, a first guess of σ^2, say σ_0^2, is made. Depending on the context, this guess could for instance stem from variographic tools, or from a previous iteration in the case of a sequential design of experiments. Based on τ^2 and σ_0^2, an approximate formula –analogue to eq. (9)– is proposed for the optimal variance as a function of the range:

$$\widetilde{\sigma^{2*}}(\theta) := \frac{1}{N} (\mathbf{g} - \widehat{\mu}(\theta)\mathbf{1})^T \left(R(\theta) + \frac{\tau^2}{\sigma_0^2} I \right)^{-1} (\mathbf{g} - \widehat{\mu}(\theta)\mathbf{1}), \tag{11}$$

where θ is finally tuned by optimizing the following approximate concentrated likelihood:

$$\widetilde{l_c}(\theta; \mathbf{g}) := l(\widetilde{\sigma^{2*}}(\theta), \theta, \tau^2; \mathbf{g}) \tag{12}$$

4. Sequential search driven by proxy-based Kriging

The Kriging model presented in the previous section allows to calculate the Kriging mean $m(\mathbf{x}_i)$ and variance $s^2(\mathbf{x}_i)$ for predicting the (transformed) misfit $g(\mathbf{x}_i)$ for any candidate model $\mathbf{x}_i$. We want then to use that information to select the candidate models on which the accurate numerical model will be executed during the search procedure in order to identify the ones with the lowest misfits. For that purpose, we propose to use a variant of the *Expected Improvement* (EI) criterion, meant to spend more time exploring the basins of optima than the genuine EI.

By definition, EI is intended to point towards promising points, but also to foster space exploration. Hence, in EI algorithms like EGO [20], a typical behavior when evaluating the objective function at a good point (i.e. at a point becoming the current best) is to spend some additional iterations in its neighborhood, and then to get attracted by unexplored regions with higher Kriging variances. This can be explained by coming back to EI's formal definition. Let us denote by $g(\mathbf{X}_n)$ the vector of (transformed misfit) observations after n accurate evaluations of the misfit function, $\min(g(\mathbf{X}_n))$ is the minimum value of the misfit found so far. The aim is now to find a location $\mathbf{x}$ in the high dimensional parameter space E such that there is a high chance to find a smaller value of the misfit. Let us remind the reader that the misfit is modeled as a Gaussian Process $(G_\mathbf{x})_{\mathbf{x}\in E}$, one can then express the possible improvement (it is a random variable) in any location of E as the difference between the current minimum and the possible value of the misfit $\min(G_{\mathbf{X}_n}) - G_\mathbf{x}$, of course only positive values must be taken into account since we are not interested in regions with worse misfit, the improvement is therefore equal to $\max(\min(G_{\mathbf{X}_n}) - G_\mathbf{x}, 0) = (\min(G_{\mathbf{X}_n}) - G_\mathbf{x})^+$. The EI criterion for a candidate model $\mathbf{x}$ then writes as the expectation of the improvement conditional to $g(\mathbf{X}_n)$:

$$EI(\mathbf{x}) : = \mathbb{E}\left[(\min(G_{\mathbf{X}_n}) - G_\mathbf{x})^+ | G_{\mathbf{X}_n} = g(\mathbf{X}_n)\right] \tag{13}$$

where conditioning on the event $G_{\mathbf{X}_n} = g(\mathbf{X}_n)$ turns $\min(G_{\mathbf{X}_n})$ into $\min(g(\mathbf{X}_n))$, and leads to the well-known Gaussian conditional distribution for $G_{\mathbf{x}}$:

$$\mathcal{L}(G_{\mathbf{x}}|G_{\mathbf{X}_n} = g(\mathbf{X}_n)) = \mathcal{N}(m(\mathbf{x}), s^2(\mathbf{x})) \tag{14}$$

Owing to this convenient property, the EI criterion offers the advantage of being analytically tractable (see [20]). Noting $T = \min(g(\mathbf{X}_n))$ and $f_{\mathcal{N}(m(\mathbf{x}),s^2(\mathbf{x}))}$ for the density of the $\mathcal{N}(m(\mathbf{x}), s^2(\mathbf{x}))$ distribution, we have indeed:

$$\begin{aligned} EI(\mathbf{x}) &= \int_{-\infty}^{T} (T-u) f_{\mathcal{N}(m(\mathbf{x}),s^2(\mathbf{x}))}(u) du \\ &= (T - m(\mathbf{x}))\Phi\left(\frac{T - m(\mathbf{x})}{s(\mathbf{x})}\right) + s(\mathbf{x})\phi\left(\frac{T - m(\mathbf{x})}{s(\mathbf{x})}\right), \end{aligned} \tag{15}$$

where Φ and ϕ stand for the cumulative distribution function and the probability distribution function of the standard Gaussian distribution, respectively. Here we propose a variant of EI meant to put more emphasis on the exploration of basins of minimum while remaining tractable. Indeed, the aim in our motivating applications is not only to find the global minimizer(s) of g as quickly as possible, but also to find *a representative subset* of inputs leading to a response curve close to the reference, i.e. to a small misfit. The proposed trick to lower the repulsion effect of current best points is to replace $\min(g(\mathbf{X}_n))$ by a quantile of $g(\mathbf{X}_n)$ in the definition of EI. Calling α the level of this quantile, we denote

$$EI_\alpha(\mathbf{x}) = (q_\alpha - m(\mathbf{x}))\Phi\left(\frac{q_\alpha - m(\mathbf{x})}{s(\mathbf{x})}\right) + s(\mathbf{x})\phi\left(\frac{q_\alpha - m(\mathbf{x})}{s(\mathbf{x})}\right) \tag{16}$$

where $q_\alpha = q_\alpha(\mathbf{X}_n)$ is the empirical $\alpha\%$-quantile of the sample of misfits $\{g(\mathbf{x}_{i_j}), 1 \leq j \leq n\}$. Varying α allows tuning the criterion from normally explorative to very local. Indeed, when $\alpha = 0$, $q_{\alpha,n}$ coincides with the minimum of $g(\mathbf{X}_n)$, so that $EI_0 \equiv EI$. However, when tuning α to a strictly positive value (obviously smaller than 1), the tendency of EI to vanish near the observation points disappears. To prevent the algorithm from resampling at already explored points, we exclude them from the search. However, we are interested in points very close to the already explored points in terms of the proposed kernel, since they have similar proxy responses but may be very different in terms of inputs. Different values of α will be investigated in the application section, where the benefit of taking $\alpha > 0$ will be illustrated.

5. Illustration of the method through a case study

In order to test the proposed approach, we consider a relatively simple but realistic example. It consists of a two-dimensional solute transport problem. The geology is based on an aquifer analogue in a glacio-fluvial environment that has been mapped in detail in the Herten site by Bayer et al. [33].

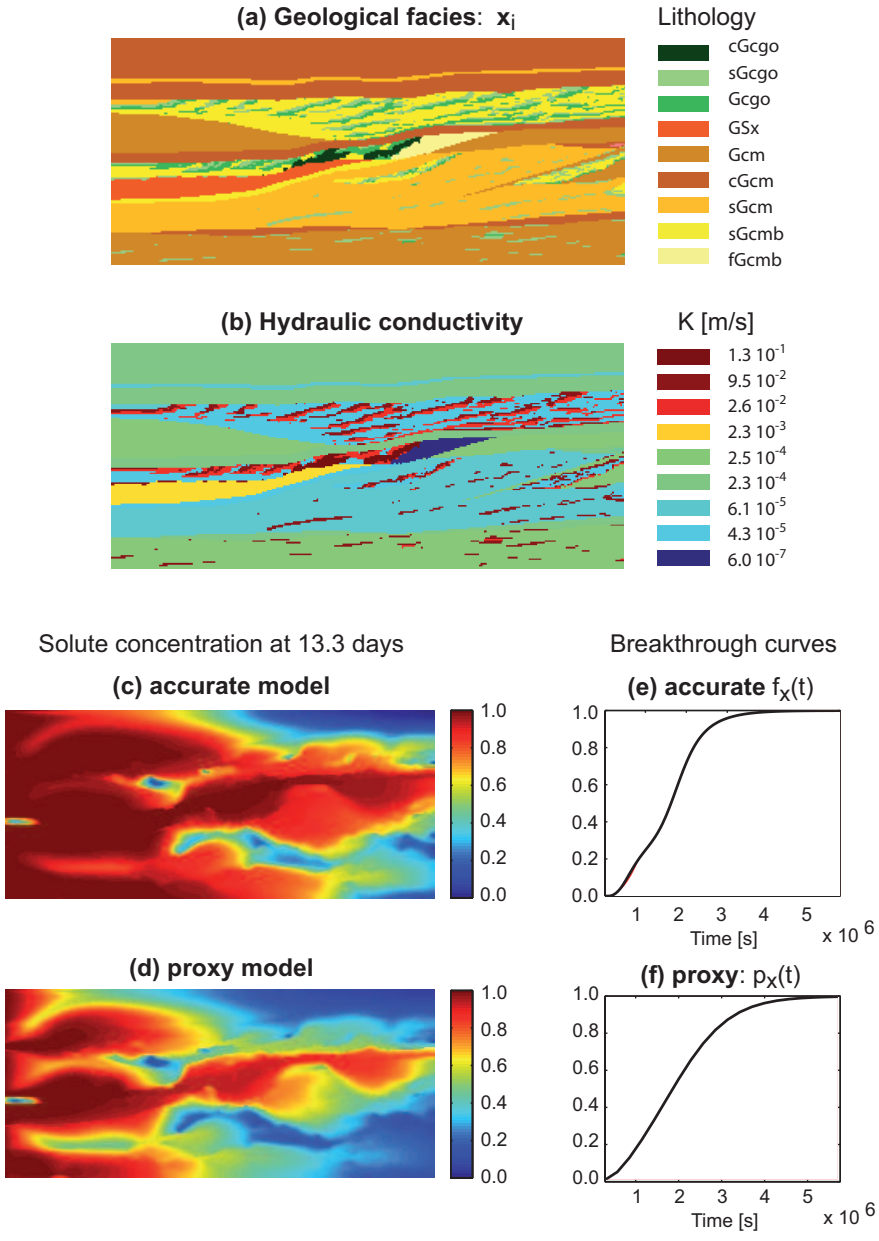

Figure 4: Illustration of the hydrogeological problem.

5.1. Geological facies simulations

To start, 1000 stochastic vertical sections of geological media $\mathbf{x}_i$ have been generated using the Direct Sampling (DS) multiple-point statistics algorithm [30] with one of the geological sections at Herten as training image [33]. The grid has a size of 320 by 140 pixels and covers an area of 16m by 7m. A few realizations are represented in Figs. 4(a) and 5(a). The realizations are constrained by a secondary variable (describing the main stratification) in the training image and in the simulations to ensure that the main sedimentary structures observed at the site are reproduced, following the approach used in 3D by Comunian et al. [34]. The parameters that were used for the DS method are: a search neighborhood of 20 cells on each axis, a maximal number of neighboring nodes of 15, a distance threshold of 0.01, and a maximal scan fraction of 0.5.

The ensemble of those geological models constitutes a sample of the prior distribution of the geological fields that are expected to occur in this environment. Fig. 5(a) displays 9 of those realizations, in which the variability between representations is present only at small scales within the main sedimentary bodies. The large scale structures are identical in all simulations.

5.2. Flow and transport simulations

The spatial discretization for the flow and transport problem is kept identical to the grid used for the geological simulations. The boundary conditions and parameters are summarized in Table 1. A constant value of the hydraulic conductivity is assigned to each facies (Fig. 4(b)) according to the mean values obtained from laboratory experiments and described by Bayer et al. [33]. For the sake of simplicity, the porosity is considered homogeneous over all facies. The flow is uniform from left to right and in steady-state. A constant head is prescribed on the left (0.1m) and right boundaries (0m). The upper and lower boundaries are no flow boundaries. The initial distribution of the solute concentration is set to zero everywhere in the domain. A fixed concentration of 1 is prescribed on the left boundary. The advective-dispersive-diffusive transport is solved in transient regime by using a finite volume matlab toolbox provided by I. Lunati [35, 36]. Figure 4(c) shows the map of

the solute concentration for the realization shown in Fig. 4(a) after 13.3 days of simulations. On the right boundary, the solute fluxes are integrated to compute the breakthrough curve $f_{\mathbf{x}}(t)$ representing the mean concentration at the outlet versus time (Fig. 4(e)).

Parameter	Value
Porosity	0.35
Molecular diffusion	$4.0 \times 10^{-9}\, m/s$
Longitudinal dispersivity (along x axis)	$0.1\, m$
Transversely dispersivity (along z axis)	$0.01\, m$
Total simulation time	$1.44 \times 10^{7}\, s$
Time steps length	$1.44 \times 10^{4}\, s$

Table 1: Parameter values for the solute transport model

Despite the apparent small variability in the geological structure discussed above, a wide range of tracer breakthrough responses are obtained on the prior ensemble (Fig.5(b)). This illustrates the importance of the internal heterogeneity of the high permeability features within the main sedimentary layers.

5.3. Two different proxies

A good proxy is faster than the accurate numerical model and allows to distinguish models that have similar or different responses in terms of tracer breakthrough. Such a proxy is generally not expected to provide an accurate simulation of the breakthrough or of solute concentration states. It should simply be a fast approximation allowing to discriminate models.

In this paper, we considered two different proxies and check their performances and reliability. The first one, $p^1_{\mathbf{x}}(t)$, is based on simplified physics. We use the same solver and the same spatial and temporal resolution as for the accurate model based on the full physics, but we disregard diffusion and dispersion effects. The numerical simulation thereby only accounts for advection and numerical dispersion phenomena. The second proxy, $p^2_{\mathbf{x}}(t)$, is based on simply coarsening the time discretization of the accurate model. The number of

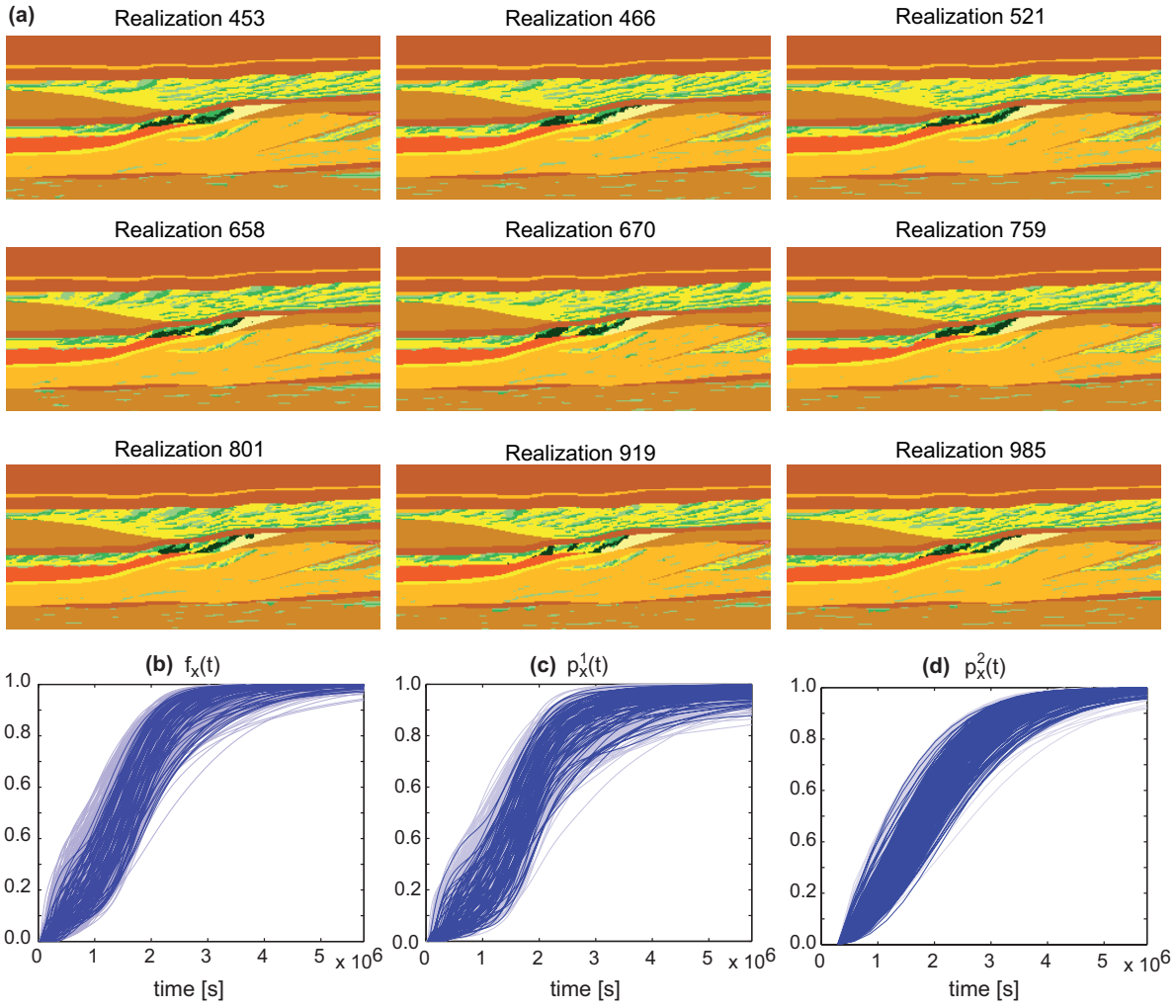

Figure 5: (a) 9 realizations of the lithofacies. Because all the simulations are constrained by the large scale structure data, only the internal architecture within the main layers is displaying some variability between the simulations. (b) Ensemble of the breakthrough curves obtained with the accurate numerical model and the two proxies (c and d) for the 1000 models. To make the figure more readable, some breakthrough curves are represented in light gray color.

time steps is reduced; their duration is increased to $2.88 \times 10^5\ s$ (i.e. a division by 20 of the number of time steps).

The breakthrough curves computed with the two proxies are displayed in Figs. 5(c) and 5(d). The first proxy gives breakthrough curves whose general shape resemble more the accurate model than the second proxy: some of the curves display a sigmoidal shape like the fine scale solution. The second proxy results in breakthrough curves that are more regular. For this proxy, the first arrivals of the tracer are almost identical for all geological models because of the coarse temporal resolution. The responses for $p^2_{\mathbf{x}}(t)$ present some variability, but less than $f_x(t)$ and the first proxy. For both proxies, the computational time is reduced by a factor of about 20. The accurate numerical solution takes about 7.5 minutes on a PC, while the two proxies run in about 20 seconds each.

5.4. Results

Let us now apply our Kriging model to the problem of predicting the transformed misfit between the breakthrough curves of a given reference and the responses associated with the 1000 candidate geological media. The proxy used here is $p^1_x(t)$, the one with simplified physics. Here we arbitrarily choose one of the actual response curves (the realization with index 800) for illustration purposes. Note that more general results will be presented in section 6, where statistics will be derived based on 100 randomly chosen reference curves.

Among the 1000 considered inputs, 50 are chosen based on a clustering technique using proxy-induced distance (Fig. 6(a)), in the flavor of Scheidt and Caers [23]'s approach. The actual response curves are calculated by using the accurate numerical model with the latter inputs, and the 50 corresponding values of misfit to the reference curve are calculated and stored in a vector, denoted by $g^\circ(X_{50})$ or $\mathbf{g}^\circ$, as in section 3.

As shown on Figure 7, a transformation is used to make the data misfits closer to Gaussian. For simplicity, we restrict the transformation to be a power transform, $g = (g^\circ)^a$. The ad hoc approach proposed here to determine the coefficient of this transform is to set the skewness of the transformed sample equal to zero. As will be presented in more detail in section 6 (performance assessment), such transform significantly improves the predictivity

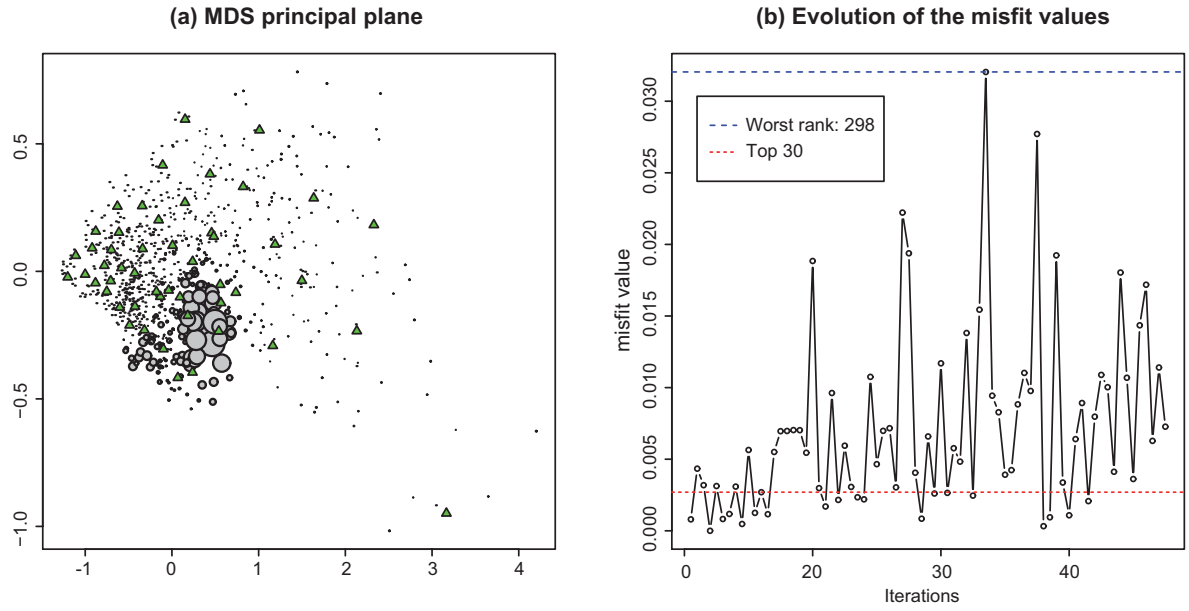

Figure 6: (a) Every point in the MDS space represents a parameter field. The triangles indicate the models that were selected by the K-means algorithm for the initial design of experiments, and the radius of the circles are proportional to the EI_α criterion; (b) Monitoring of the misfit values obtained for the parameter fields sequentially chosen by the ProKSI algorithm.

of the Kriging model, as well as the performances of the inversion algorithm proposed in the next section.

In a second step, we estimate the kernel parameters by maximum likelihood (MLE) based on the transformed sample of fits. We can see in Fig. 8 that the optimal value of θ is very clearly defined since the log-likelihood curve has a large curvature at its minimum value.

The quality of the resulting Kriging estimates is then evaluated: we first use a standard cross validation technique on the 50 samples used to infer the Kriging model (Fig. 9(a)) and then extend the comparison to an external validation on the complete ensemble of 1000 values (Fig. 9(b)). In both cases, the predicted values obtained by Kriging are in good agreement with the true values; the regression line of predicted versus actual values has in intercept $B0$ close to zero and a slope $B1$ close to 1 (Fig. 9), indicating that the Kriging predictions are not notoriously biased. Furthermore, one can see that the leave-on-out errors of (a) give a reasonable estimate of the prediction errors observed a posteriori on

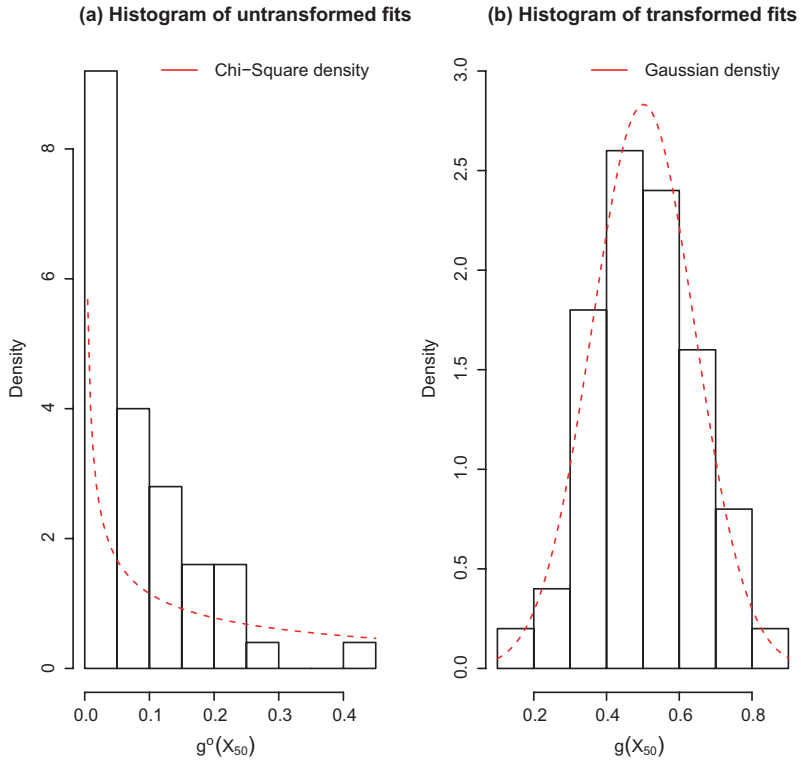

Figure 7: Samples of untransformed (left) and transformed (right) misfit values obtained at a 50-point initial design of experiments in the case of a proxy with simplified physics. The histogram of the untransformed sample is closer to a chi-square distribution, whereas the one obtained by a power transformation, although remaining positive, is much more similar to the one of a Gaussian sample. The exponent used in the power transformation ($a \approx 0.24$ here) is obtained by setting the skewness of the transformed sample to 0.

the exhaustive validation set.

6. Performance assessment

The good results obtained in the illustrating example above (Fig. 8) are of course conditioned by the chosen reference breakthrough curve f_{ref} (here with index 800) and do not constitute a sufficient basis to appraise the ProKSI algorithm. Furthermore, the method is proxy-dependent, and it would make sense to test the sensitivity of the performances to both an improvement or a degradation in the proxy. In this section, we propose a more systematic benchmarking of the algorithm's performances by analyzing the results obtained with 100 different f_{ref}'s, and for three different proxies, with a comparison to Monte Carlo

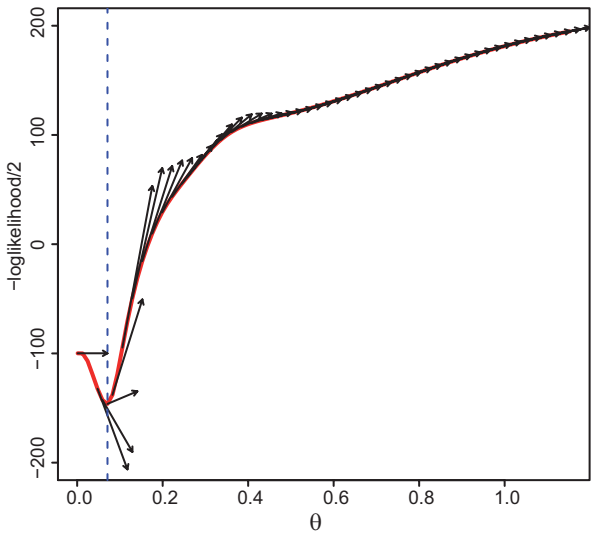

Figure 8: Identification of θ by maximizing the concentrated log-likelihood function. The large curvature at the minimum indicates a well-identified parameter value. The arrows stand for the gradients of the log-likelihood function, which have been calculated analytically, and are used within the optimization procedure.

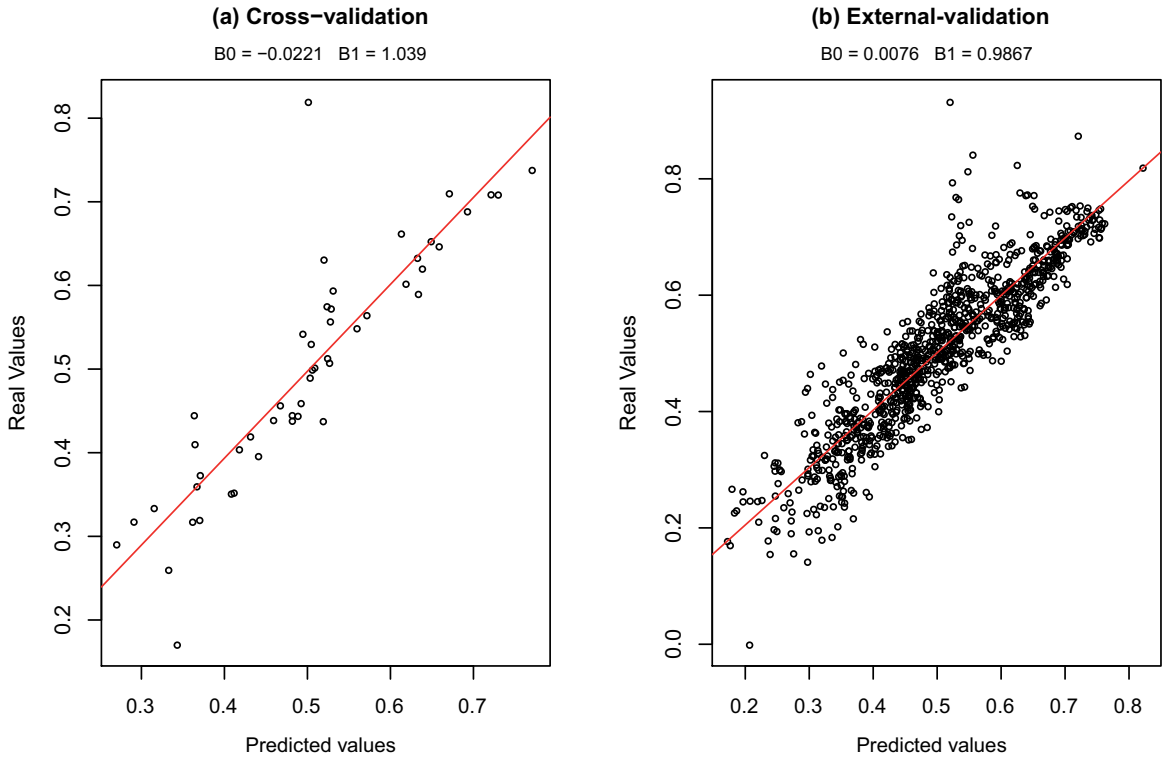

Figure 9: (a) Cross validation and (b) External validation

random search in the case of the worse proxy. In that last situation, we will use a completely inadequate proxy model to test the robustness of the method. Furthermore, the effect of the power transform applied to the misfit function, as well as the effect of the replacement of the minimum by a quantile in the EI criterion are investigated. Before giving more details about the benchmark and the obtained results, let us first present the main performance evaluation metrics.

6.1. Performance evaluation metrics

EM1: current best model's rank. One of the most natural way of evaluating an optimization method consists in monitoring the evolution of the misfit as a function of the number of iterations (Fig. 6(b)). One can also plot the smallest misfit value achieved so far as function of the number of iterations. However, the curve obtained for such a metric would have a scale (on the y-axis) depending on the considered f_{ref}, which would prevent us from making comparisons between different tests. As a consequence, we choose to focus on the evolution of the rank of the current best model among the 1000 candidates. This rank would normally be unknown but here we can compute it because we evaluate the true misfit for all the candidate models (even those which are not selected by the ProKSI algorithm) in order to be able to test the efficiency of the method. Because, we then use multiple references and because we repeat the numerical experiment, we can then plot some statistics of the rank as a function of the number of iterations (Fig. 10(a)).

EM2: number of evaluated models from the top 30. The first metric (EM1) focuses on the capacity of the method to find at least one model with a low misfit value, but not on its ability to explore the set of models with low misfit values. EM2 is meant to be a complement to EM1, by measuring the number of models of the top 30 (i.e. the 3% best models in terms of misfit value) evaluated along the algorithm. Though rather arbitrary, EM2 gives a good picture of the algorithm's tendency to explore the possible multiple peaks of the posterior distribution of models. Again, the statistics of EM2 are plotted as a function of the number of iterations (Fig. 10(b)).

EM3: probability that random search outperforms the proposed algorithm. It is expected that an elaborated algorithm like ProKSI (relying on a metamodel) performs better than random search, and at least not much worse in cases where the proxy is misspecified. The metric EM2 is well-adapted to base a comparison of ProKSI to a naive Monte Carlo (MC) algorithm, since the law of the number of points visited in the top 30 can be analytically derived for the case of a random search (this number then follows a hyper-geometric distribution). EM3 is a curve giving at each iteration of ProKSI the probability that an MC algorithm finds more points in the top 30.

6.2. Benchmark: design and implementation

6.2.1. Design of the benchmark

The aim of the benchmark was to assess the global performances of the ProKSI algorithm on the considered case study with the following specific questions in mind. **How sensitive are the performances to: (Q1) the chosen proxy, (Q2) the value of the quantile α, and (Q3) the normalizing transform of the misfit values?**

Consequently, we ran replications of the algorithm (by varying the reference curve) with different proxies, with or without power transform of the misfit function, and with different values of α. In order to have results based on solid statistical analysis, rather than on an arbitrary set of examples with a potentially low generalization power, we ran the ProKSI algorithm 100 times for each configuration (i.e. for each considered (proxy, transform, α) combination). For each proxy considered (p^1, p^2, and a third mismatched one described below), 50 models are chosen by Scheidt and Caers clustering technique based on Multi-Dimensional Scaling, and 100 f_{ref} are randomly chosen among the 950 remaining models. Then, for any given configuration (in terms of transform and/or α value), 75 iterations of the ProKSI algorithm are run for the 100 chosen f_{ref}. The results are visualized in terms of box-plot sequences representing the statistical distributions of 100 values for the considered evaluation metric, evolving over the 75 iterations. Finally, for EM3, one sequence of 75 probabilities that a Monte Carlo algorithm would lead to more points in the top 30 than the proposed approach (one probability per iteration) can be produced for each replicate. Then,

in the same way as previously, one may sum up the results for any given configuration by representing sequences of box-plots.

6.2.2. Implementation of the benchmark

All the benchmark algorithm runs and the performance evaluation calculations were done using the open source statistical software R, based on the numerical simulation results obtained for the 1000 multiple-statistics simulations (see implementation details in section 5). The R code, gathered in form of a package (*ProKSI*, forthcoming on the *Comprehensive R Archive Network*), was called for each task of the following loop, forming the basic brick of the benchmark for any fixed configuration:

Algorithm 1 Testing procedure for a proxy with a given algorithm configuration

1: **Choose** the initial design of experiment (**50** points using Scheidt et Caers approach).
2: **Choose 100** different simulations among the 950 remaining points.
3: **for** $i = 1$ to $i = 100$ **do**
4: **Run 75** iterations of the algorithm on the i^{th} reference.
5: **Evaluate** the 3 EM's for each iteration of the i^{th} run.
6: **end for**

6.3. Results

The first benchmark results, displayed on Figure 10, deal with the performances on the ProKSI algorithm when applied to our test-case with proxy 1, and default settings concerning the normalizing transform and the EI variant (power transformation done, and $\alpha = 0.15$). Figure 10(a) represents the evolution of the statistics (box-plot) of EM1 over the 100 replicates, along the 75 iterations of the algorithm. We can see here that in 42 iterations, the actual best model has been found for more than 50% of the replications. Figure 10(b), the exploration performances are investigated in terms of EM2; it is found here that 15 models among the 30 best ones (out of 1000) have been evaluated in median after 75 iterations of the algorithm. In total, these results show both how the proposed Kriging

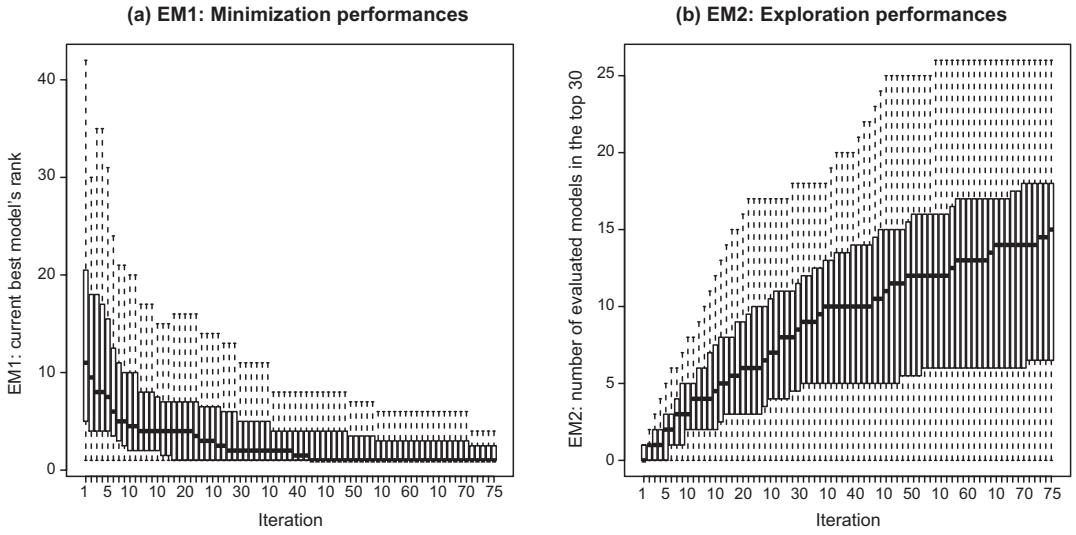

Figure 10: Performances of the ProKSI algorithm (based on proxy 1) with a power transform of the misfit. (a) box-plot of the EM1 criteria over the 100 replicates of the numerical experiment. (b) box-plot of the EM2 criteria.

metamodel helps reaching a fast convergence, and that ProKSI achieves a rather satisfying exploration of the set of best models in a limited number of iterations.

6.3.1. Effect of the misfit transformation on the algorithm performances

Figure 11 represents the performances (in terms of EM1 and EM2) obtained by applying the ProKSI algorithm to our case study with default settings concerning the EI criterion ($\alpha = 0.15$) but without normalizing power transform for the misfit function.

The results appear to be clearly inferior to the ones obtained with the transformation: here, even after the 75 iterations, the median rank of the best evaluated model is strictly above 1, which expresses a significantly slower convergence of ProKSI as with the transformed misfits. Similarly, the number of models forming the top 30 evaluated along the algorithm stagnates around 8 in median after the 75 iterations. The normalizing transform has thus clearly a positive effect on the efficiency of the algorithm, both in terms of fast convergence to the best model, and in terms of global exploration of the nearly optimal models.

However, as illustrated on figure 12, the results in terms of EM2 are still good enough to outperform a pure random search (upper right graphic). On the lower graphic, the evolution

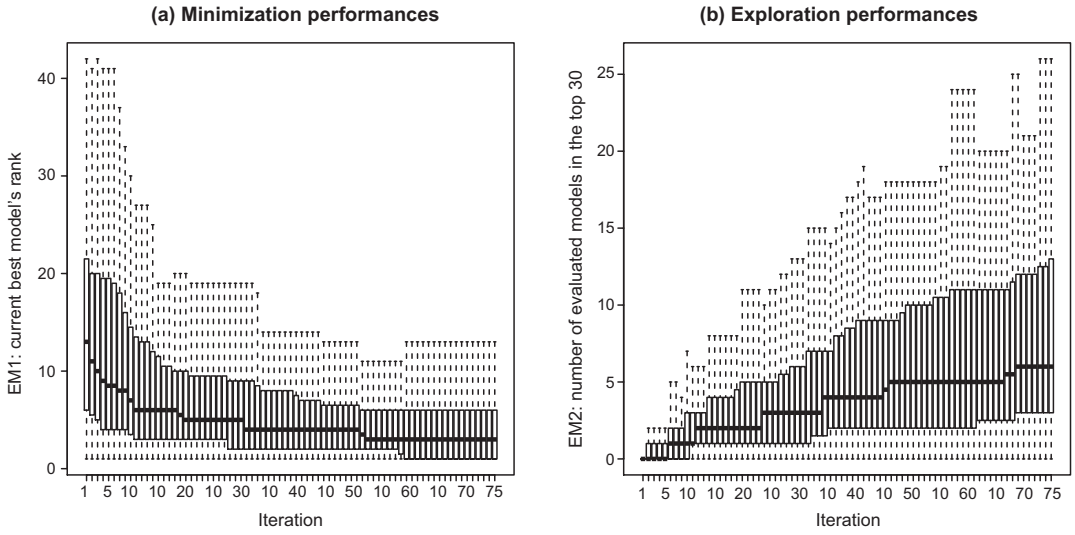

Figure 11: Performances of the ProKSI algorithm (based on proxy 1) without power transform of the misfit. (a) box-plot of the EM1 criteria over the 100 replicates of the numerical experiment. (b) box-plot of the EM2 criteria.

of the median rank for the models evaluated by ProKSI with or without misfit transform finally illustrate the trend of the algorithm with misfit transform to spend more time in the nearly optimal regions.

6.3.2. Effect of an improved proxy on the algorithm performances

Let us now present the results obtained when using the second proxy, with default settings. The most striking result when looking at figure 13 is the impressively fast convergence of the algorithm in terms of EM1 criterion. Indeed, in 7 iterations, the minimizing model has been found in all cases (100 replicates) considered. ProKSI successfully relies here on the information given by proxy 2 for uncovering the best point, only based on slightly more than the misfit values for the set of 50 initial models. What seems really outstanding in that case is that such a result is uniformly obtained for the 100 reference curves. To milden this success a bit, let us remark that the performances in terms of exploration are comparable to the first proxy, that is one half of the top 30 models were evaluated in median after termination.

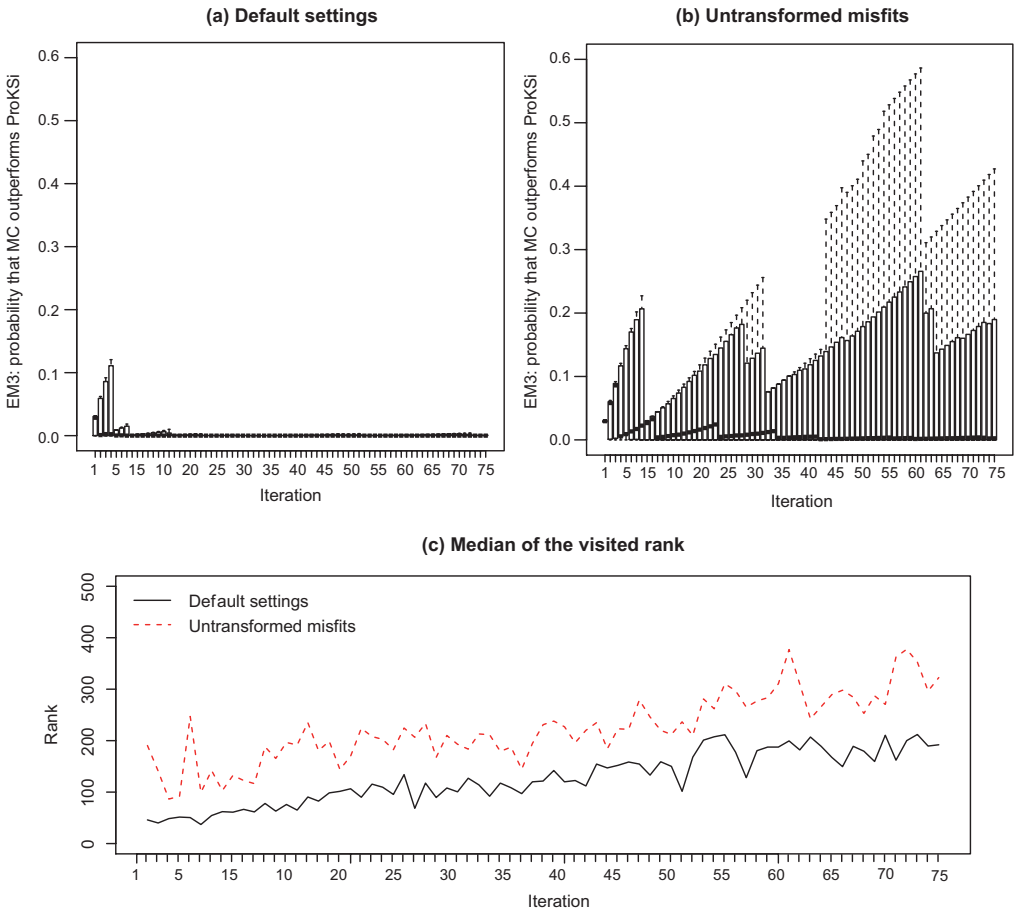

Figure 12: Effect of the misfit transformation on the performances of the ProKSI algorithm in terms of its superiority with respect to a Monte-Carlo search and median rank of the evaluated models.

6.3.3. Effect of the α parameter (from EI_α) on the algorithm performances

We investigate here the effect of the parameter α, tuning the quantile level in the proposed generalization of EI, on the performances of the algorithm. We obtained very different results for the two proxy. Indeed, the performances of ProKSI were not very sensitive to α when using the first proxy, so that we do not discuss this case here, and refer the interested reader to the appendix for more detail. However, α was found to be strongly influencing the algorithm's performances when using the second proxy, as illustrated on Figure 14.

It is indeed observed on Figure 14 (a and b) that using proxy with the standard EI criterion ($\alpha = 0$) is less efficient compared to the considered default value $\alpha = 0.15$: even though the

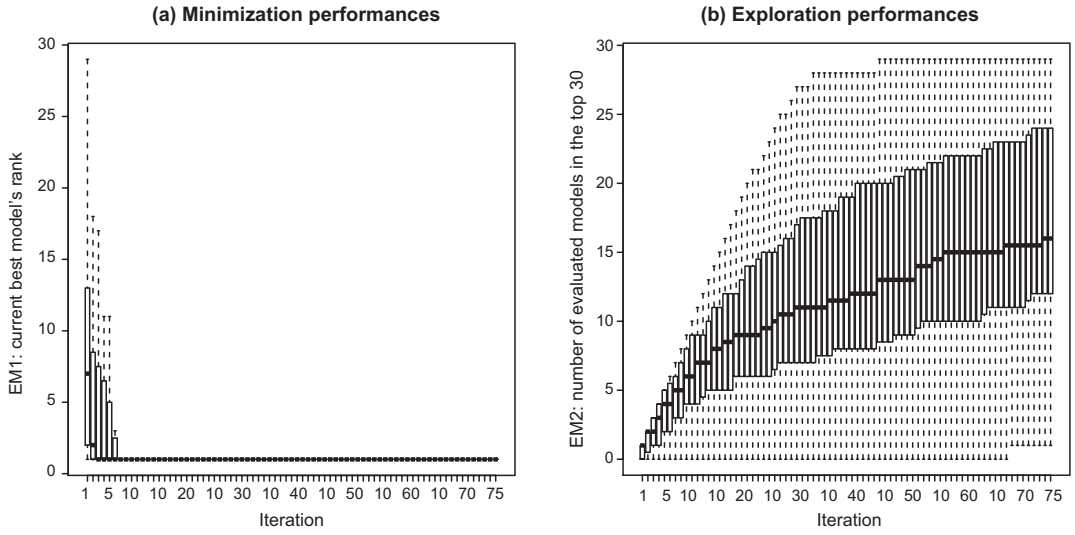

Figure 13: Performances of the ProKSI algorithm (based on proxy 2) with default settings.

algorithm convergence to the minimum is always comparably fast, the exploration performances are strongly affected by this change of criterion (median number of points in the top 30 after termination decreased from 15 to 10). On the other hand, increasing *alpha* to 0.6 was found to greatly improve the results in terms of exploration (again, without affecting the minimization performances, see 14 (c)) since the median number of points in the top 30 jumped to 25, as can be seen on 14 (d). To sum up, introducing this parameter α was found beneficial for the exploratoriness of the algorithm. Its optimal tuning is of course problem-dependent. The rather arbitrary default value $\alpha = 0.15$ chosen here gave improved results in both cases considered, even though better performances could be reached by using a larger α value in the case of the second proxy.

6.3.4. Effect of a non-informative proxy on the algorithm performances

Finally, we propose to test the performances of ProKSI when using a completely inadequate proxy model. The idea is to see if the algorithm remains consistently applicable when the simplified model is poorly (or not at all) informative, and how using ProKSI in such degraded conditions would perform compared to a naive Monte Carlo search. In order to emulate a non-informative proxy, we started from proxy 1, and randomly permuted the 1000 indices. We then ran the ProKSI algorithm with this "mismatched" proxy, and compared them to

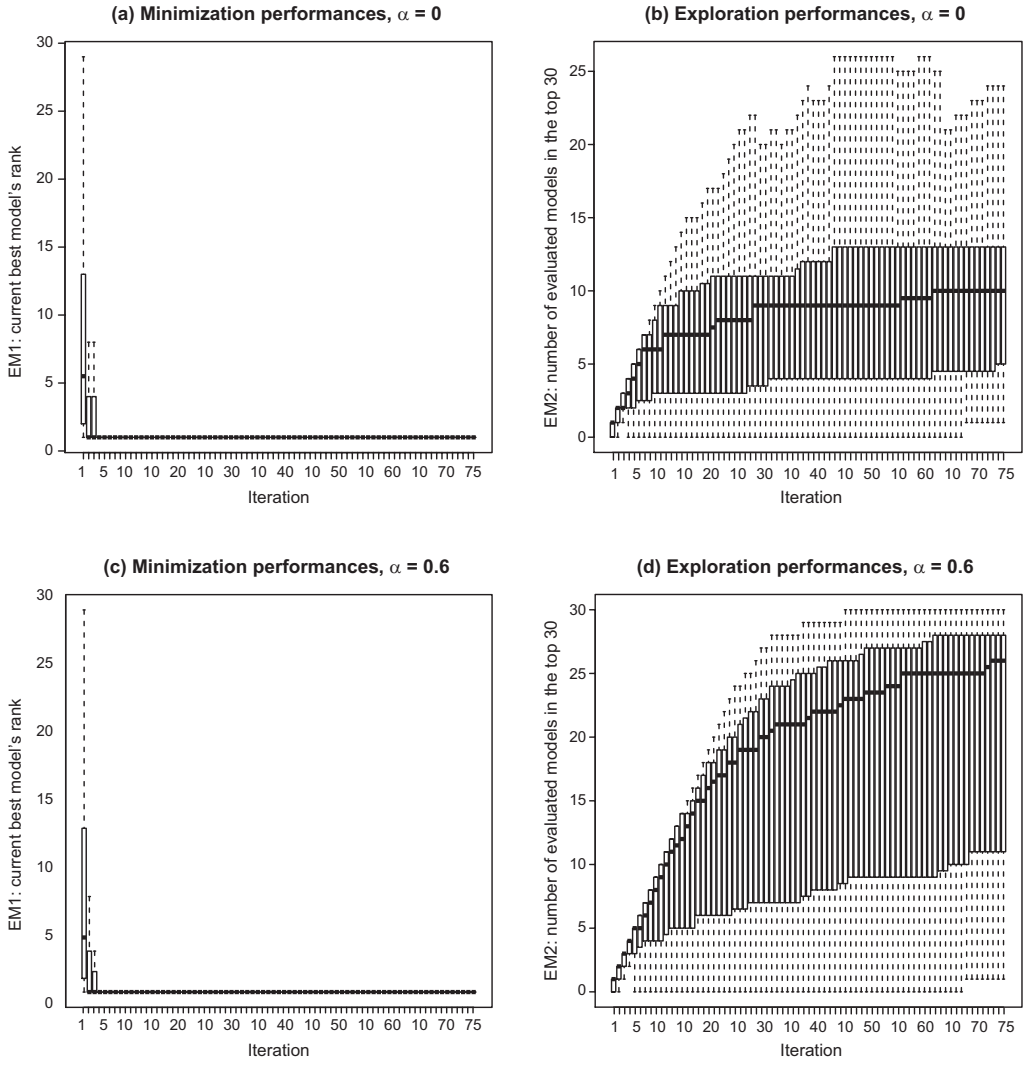

Figure 14: Effect of the α parameter on the performances when using the second proxy

trajectories obtained by Monte Carlo (the whole replicated for the 100 reference curves). As illustrated on Figure 15, the performances of ProKSI with "mismatched" proxy are comparable to those of Monte Carlo in terms of exploration, and remain significantly better in minimization. The algorithm hence appears reasonably robust to a proxy mispecification, while being potentially very efficient for well-chosen proxies, as seen previously.

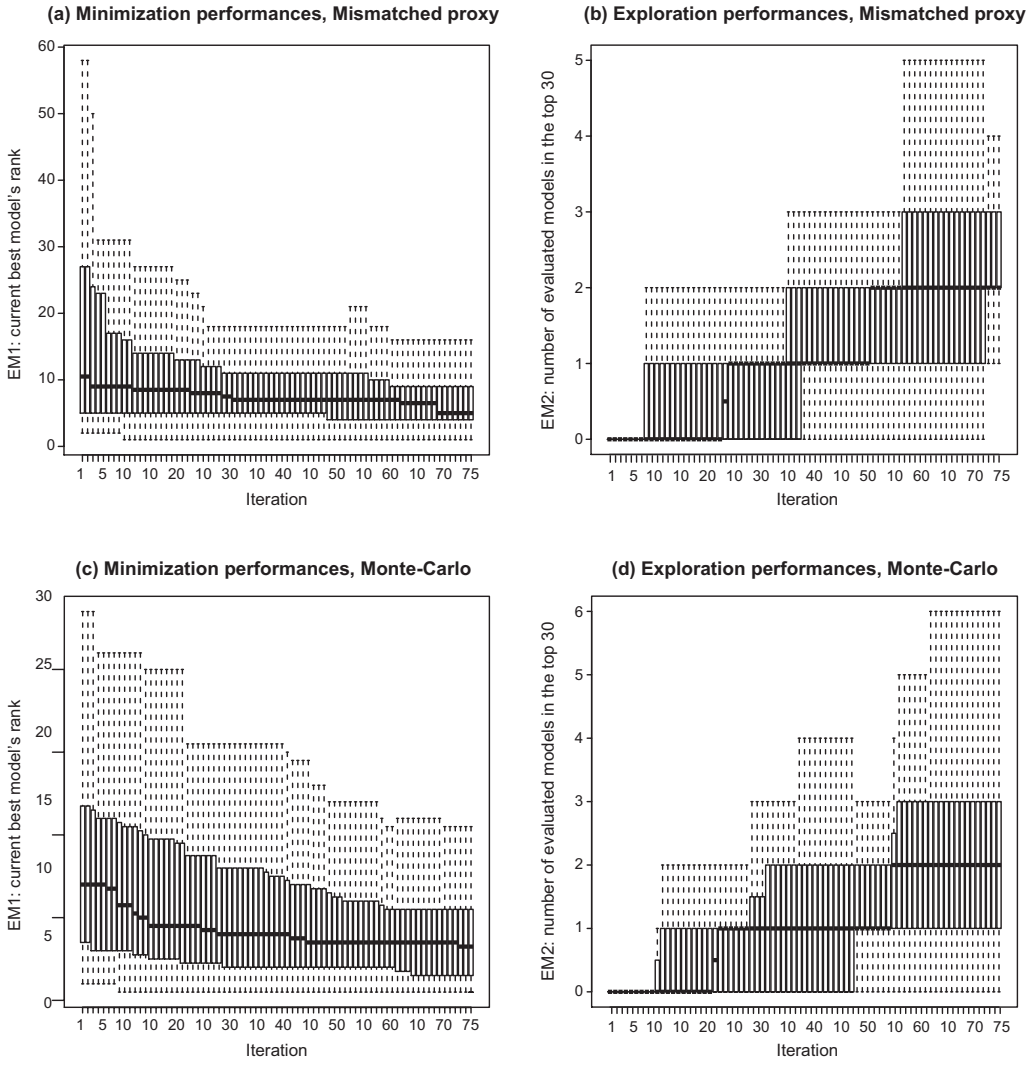

Figure 15: Effect of a non-informative proxy on the performances.

7. Conclusion

Handling complex solvers requiring heavy computational load while representing uncertainty is often contradictory. Accurate complex solvers are too computationally demanding to be used in the general framework of a Monte Carlo approach and analytical propagation of uncertainty is often intractable. Resolving this issue is an important research topic both from a theoretical perspective and for a wide range of applications [37, e.g.], including hydrogeology.

In this paper, we propose a contribution which consists in coupling a complex model (the

accurate model), a simple model (the proxy), and a statistical metamodel. The statistical metamodel is used to link the results of the proxy with those of the accurate model. More precisely, this is achieved by developing a specific covariance kernel accounting for the difference in responses from the proxy models and allowing to predict statistically the response of the accurate model using Kriging. One of the strengths of this idea is that the use of the distance between proxy responses permits to drastically reduce the dimension of the Kriging problem and allow an efficient inference of the parameters of the covariance kernel. The quality of the relation between the accurate and the proxy models is also directly taken into account via the covariance kernel. In addition, the chosen covariance kernel can be tailored to the practical problem that has to be solved (through the proxy, the kernel k_F, and more), which makes the approach flexible.

In the example case study, we showed how such an approach can help in the case of an inverse problem where the prediction refers to the misfit between observations and the accurate model responses. As a first step, we propose here an iterative search algorithm. This method is an extension of previous work done by Caers and colleagues [24, 23, 7] in which we add a step based on the use the Kriging model described above to orient the search. We propose to guide the selection of a model during the search by defining a modified Expected Improvement criterion EI_α such that the algorithm will explore potentially multiple minima if they exist.

The systematic analysis of the case study showed the following results.

- When the proxy is informative, the method is extremely efficient in finding the model parameters that minimize the misfit.
- When the proxy is less informative, the method efficiency decreases but is always much better than a random search.
- The proposed modified expected improvement criteria allows both identifying the global minimum and exploring the various basins of minimum when they exist.
- The method is more efficient when the misfit are properly transformed so as to get

a close-to-Gaussian sample. This is not surprising, because otherwise the expected statistical distribution of the misfit for a given model would not be properly predicted and the value of the Expected Improvement criteria could be biased.

- The parameter α - defining the quantile of the misfit distribution below which a model is considered as an interesting candidate - allows to control the degree of exploration of the method. A low value of α will preferentially sample the regions around the global minimum and let the algorithm behave like a maximum likelihood technique. A higher value of α will sample preferentially in the whole range of areas of minimum and will be more explorative.

We consider, that the results obtained so far are very encouraging and show that the use of a Kriging technique to couple a complex and simple model will open a broad range of new perspectives. The proposed technique can already be used directly to identify rapidly maximum likelihood solutions. If one wants to obtain not only the best solution but an ensemble of models, then the selection criterion and the iterative search procedure will have to be modified in order to ensure that the final ensemble will be a representative sample of the posterior distribution. The method can also be extended in a relatively straightforward manner to allow generating new candidate models by coupling it, for example, with the Iterative Spatial Resampling method [11]. Finally, it is also very clear that this type of approach can be parallelized to improve the numerical performances [38].

Acknowledgments

The work presented in this paper is part of the *Integrated methods for stochastic ensemble aquifer modeling (ENSEMBLE)* project supported by the Swiss National Science Foundation under the contract CRSI22_122249/1. The authors want to thank I. Lunati, who provided the finite volume matlab code to solve the flow and transport equations and helped to setup the flow problem and proxy simulations, as well as N. Linde and P. Brunner for their constructive review comments.

[1] de Marsily, G., Delhomme, J.P., Delay, F., Buoro, A.. 40 years of inverse problems in hydrogeology. Comptes Rendus de l'Academie des Sciences Series IIA Earth and Planetary Science 1999;329(2):73–87.

[2] Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., Slooten, L.. Inverse problem in hydrogeology. Hydrogeology Journal 2005;13(1):206–222.

[3] Tarantola, A.. Inverse Problem Theory and Model Parameter Estimation. SIAM; 2005.

[4] Hendricks Franssen, H.J., Alcolea, A., Riva, M., Bakr, M., van der Wiel, N., Stauffer, F., et al. A comparison of seven methods for the inverse modelling of groundwater flow. application to the characterisation of well catchments. Advances in Water Resources 2009;32(6):851–872. Franssen, H. J. Hendricks Alcolea, A. Riva, M. Bakr, M. van der Wiel, N. Stauffer, F. Guadagnini, A.

[5] Oliver, D., Chen, Y.. Recent progress on reservoir history matching: a review. Computational Geosciences 2011;15:185–221.

[6] Mosegard, K., Tarantola, A.. Monte carlo sampling of solutions to inverse problems. Water Resources Research 1995;100(B7):12431–12447.

[7] Caers, J.. Modeling uncertainty in the earth sciences. Wiley-Blackwell; 2011.

[8] Vrugt, J.A., Ter Braak, C.J.F.. Dream(d): an adaptive markov chain monte carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems. Hydrol Earth Syst Sci, 2011;15:37013713.

[9] Fu, J.L., Gomez-Hernandez, J.J.. A blocking markov chain monte carlo method for inverse stochastic hydrogeological modeling. Mathematical Geosciences 2009;41(2):105–128.

[10] Alcolea, A., Renard, P.. Blocking moving window algorithm: Conditioning multiple-point simulations to hydrogeological data. Water Resources Research 2010;46.

[11] Mariethoz, G., Renard, P., Caers, J.. Bayesian inverse problem and optimization with iterative spatial resampling. Water Resources Research 2010;46(W11530).

[12] Keating, E.H., Doherty, J., Vrugt, J.A., Kang, Q.J.. Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality. Water Resources Research 2010;46(W10517).

[13] Bliznyuk, N., Ruppert, D., Shoemaker, C., Regis, R., Wild, S., Mugunthan, P.. Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation. Journal of Computational and Graphical Statistics 2008;17(2):270–294.

[14] Mugunthan, P., Shoemaker, C.A.. Assessing the impacts of parameter uncertainty for computationally expensive groundwater models. Water Resources Research 2006;42(10).

[15] Sacks, J., Welch, W., Mitchell, T., Wynn, H.. Design and analysis of computer experiments. Statistical Science 1989;4(4):409–435.

[16] Matheron, G.. Principles of geostatistics. Economic Geology 1963;58:1246–1266.

[17] Rasmussen, C., Williams, K.. Gaussian Processes for Machine Learning. M.I.T. Press; 2006.

[18] Paciorek, C.. Nonstationary gaussian processes for regression and spatial modelling. Ph.D. thesis; Carnegie Mellon University; 2003.

[19] Mockus, J.. Bayesian Approach to Global Optimization. Kluwer academic publishers; 1988.

[20] Jones, D., Schonlau, M., Welch, W.. Efficient global optimization of expensive black-box functions. Journal of Global Optimization 1998;13:455–492.

[21] Bect, J., Ginsbourger, D., Li, L., Picheny, V., Vazquez, E.. Sequential design of computer experiments for the estimation of a probability of failure. Statistics and Computing 2011;doi:10.1007/s11222-011-9241-4.

[22] Suzuki, S., Caers, J.. A distance-based prior model parameterization for constraining solutions of spatial inverse problems. Mathematical Geosciences 2008;40(4):445–469.

[23] Scheidt, C., Caers, J.. Representing spatial uncertainty using distances and kernels. Mathematical Geosciences 2009;41(4):397–419.

[24] Suzuki, S., Caumon, G., Caers, J.. Dynamic data integration for structural modeling: model screening approach using a distance-based model parameterization. Computational Geosciences 2008;12:105–119.

[25] Sambridge, M.. Geophysical inversion with a neighborhood algorithm-i: searching a parameter space. Geophys J Int 1999;138(2):479–494.

[26] Caers, J., Park, K., Scheidt, C.. Handbook of Geomathematics; vol. Part 5; chap. Modeling uncertainty of complex earth systems in metric space. Springer; 2011, p. 865–889.

[27] Kennedy, M.C., O'Hagan, A.. Predicting the output from a complex computer code when fast approximations are available. Biometrika 2000;87(1):1–13.

[28] Lodoen, O.P., Tjelmeland, H.. Bayesian calibration of hydrocarbon reservoir models using an approximate reservoir simulator in the prior specification. Statistical Modelling 2010;10(1):89–111.

[29] Doherty, J., Christensen, S.. Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. Water Resources Research 2011;47(W12534).

[30] Mariethoz, G., Renard, P., Straubhaar, J.. The direct sampling method to perform multiple-point geostatistical simulations. Water Resources Research 2010;46.

[31] Guttorp, P., Sampson, P.. Nonparametric estimation of nonstationary spatial covariance structure. Journal of the American Statistical Association 1992;87, No. 417:108–119.

[32] Santner, T., Williams, B., Notz, W.. The Design and Analysis of Computer Experiments. Springer; 2003.

[33] Bayer, P., Huggenberger, P., Renard, P., Comunian, A.. Three-dimensional high resolution fluvio-glacial aquifer analog - part 1: Field study. Journal of Hydrology 2011;405:1–9.

[34] Comunian, A., Renard, P., Straubhaar, J., Bayer, P.. Three-dimensional high resolution fluvio-glacial

aquifer analog - part 2: Geostatistical modeling. Journal of Hydrology 2011;405:10–23.

[35] Künze, R., Lunati, I.. A matlab toolbox to simulate flow through porous media. Tech. Rep.; University of Lausanne, Switzerland; 2011.

[36] Künze, R., Lunati, I.. An adaptive multiscale method for density-driven instabilities. Journal of Computational Physics 2012;:in review.

[37] Christie, M., Cliffe, A., Dawid, P., Senn, S.. Simplicity, Complexity, and Modelling. Statistics in Practice; Chichester: John Wiley & Sons, Ltd; 2011.

[38] Ginsbourger, D., Le Riche, R., Carraro, L.. Computational Intelligence in Expensive Optimization Problems; chap. "Kriging is well-suited to parallelize optimization". Studies in Evolutionary Learning and Optimization; Springer-Verlag; 2010, p. 1867–4534.

Appendix A. Proof that a p.d. kernel chained with a proxy is p.d.

Property Let E and F be two arbitrary spaces. Given a positive-semidefinite kernel k_F over $F \times F$, the kernel k_E defined by

$$k_E(\mathbf{x}, \mathbf{y}) := k_F(p(\mathbf{x}), p(\mathbf{y})) \tag{A.1}$$

is a positive-semidefinite kernel over $E \times E$ whatever the function $p : E \longrightarrow F$.

Proof. Let $n \in \mathbb{N}$, $\mathbf{x}_1, \ldots, \mathbf{x}_n \in E$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Then

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k_E(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k_F(p(\mathbf{x}_i), p(\mathbf{x}_j)) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k_F(\mathbf{y}_i, \mathbf{y}_j) \geq 0\end{aligned}$$

by using the definition of positive-definiteness applied to k_F with the points $\mathbf{y}_i := p(\mathbf{x}_i) \in F$ $(1 \leq i \leq n)$ and the coefficients $\alpha_1, \ldots, \alpha_n$ as above.

Appendix B. Supplementary figures

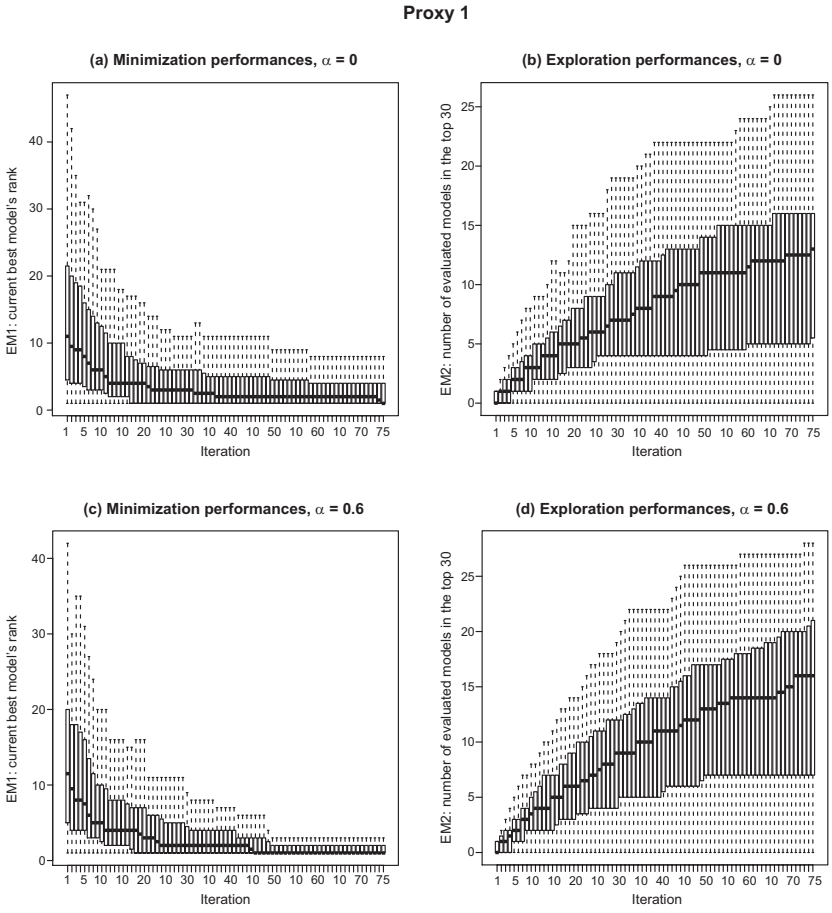

Figure B.16: Effect of α on the performances of the ProKSI algorithm for proxy 1.

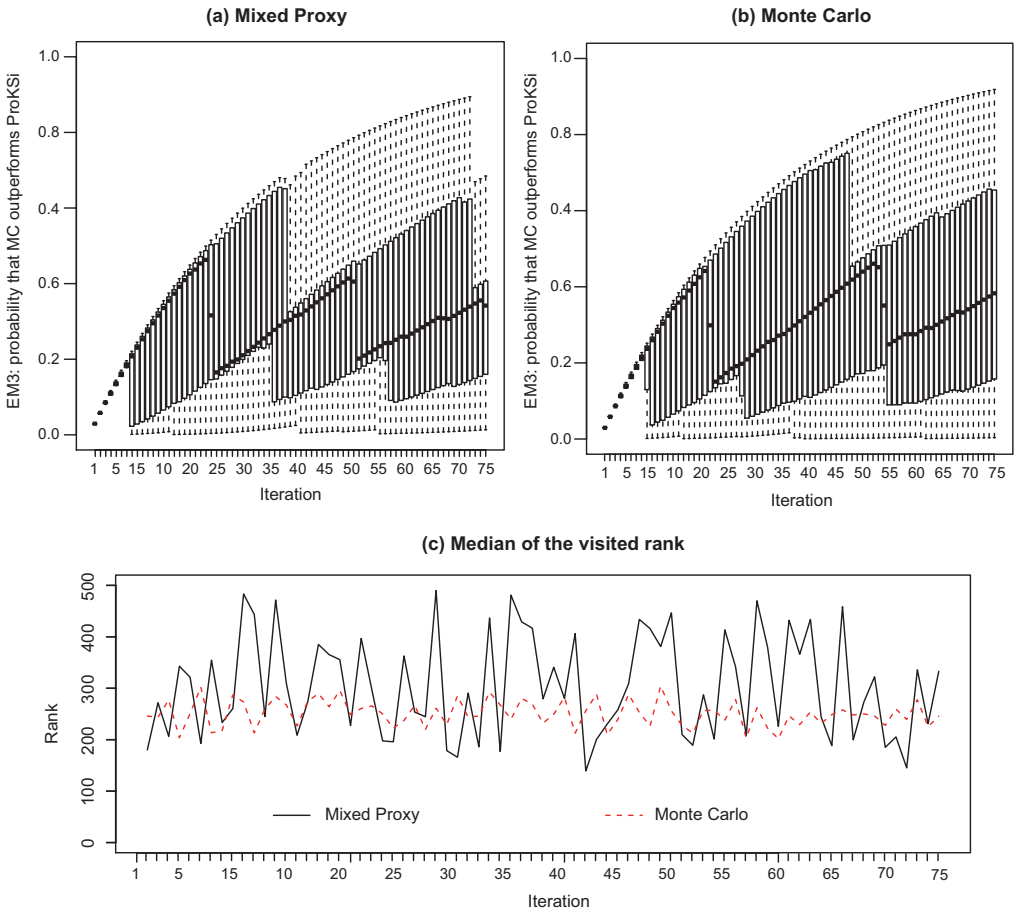

Figure B.17: Comparison of the performances of ProKSi with a purely random sampling strategy in the case of a wrong proxy.